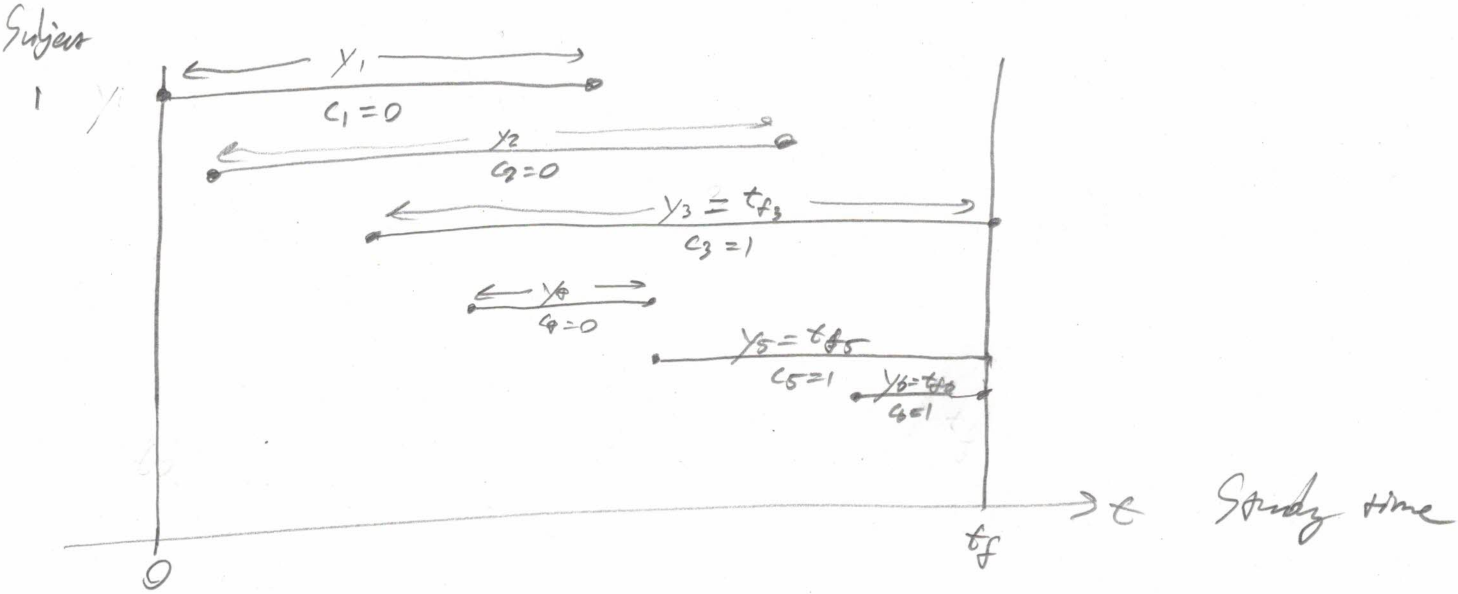
MATH 343 LEC 8

Right censoring at $t_f := \max(y)$ is simple but also not realistic. Why? Most survival settings have a max study time t_f but subjects begin at different times so..



For parametric modeling, just use t_{f_i} for $\{i : c_i = 1\}$

Why not upgrade our empirical survival function to ignore censoring i.e.

$$\hat{S}(y) = \frac{1}{n} \sum \mathbb{1}_{\substack{y_i > y \text{ if } c_i = 0 \\ y_i > t_{f_i} \text{ if } c_i = 1}}$$

or just drop censored observations?

survival biased downwards as censored measurements count as deaths

Survival biased downwards or upwards depending on the censored observations.

This is where Kaplan and Meier (1958) comes to the rescue.

Assumptions

② Random and independent censoring. Reasonable?

① Survival distr same for all subjects. ''

③ Events are recorded instantaneously so there's no confusion between censored and uncensored events

Their idea is to remove all censored observations from the "at risk pool" and not adjust the survival function at censored times

let $t_1, \dots, t_{n'}$ be the unique ordered times for actual deaths

let $s_1, \dots, s_m$ be the unique ordered times for censorings

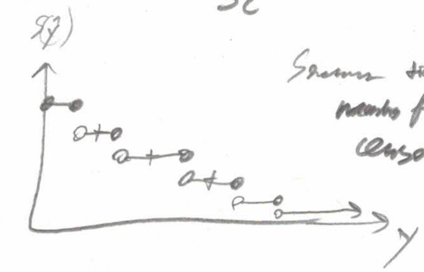
let q_i be the # of censored observations at time s_i.

in chronological order s_i's are interspersed among the t_i's.

time	d_i	q_i	n_i
$t_0 = 0$	0	0	$n_1 = n$
t_1	d_1	0	$n_2 = n_1 - d_1 - q_1$
s_1	0	q_1	N/A
t_2	d_2	0	$n_3 = n_2 - d_2$
t_3	d_3	0	$n_4 = n_3 - d_3 - q_2$
s_2	0	q_2	N/A
$\vdots$			

N/A because the $\hat{S}(t)$ function does not change at the values of s_i

Looks the same



Same w/ ticks marks for censoring

Kaplan–Meier Estimator (K-M)

$$\Rightarrow \hat{S}(y) = \prod_{\{i : y < t_i\}} \left(1 - \frac{d_i}{n_i}\right) \neq$$ the empirical survival function if there is censoring interspersed

Inference for values? It can be proven to be asymptotically normal as it can be shown to be a type of MLE

Bootstrap or....

$$\hat{S}(y) \,\dot{\sim}\, N\left(S(y), \mathrm{Var}[\hat{S}(y)]\right)$$

Many expressions for variance. Most common is

$$\mathrm{Var}[\hat{S}(y)] = \hat{S}(y)^2 \sum_{\{i : y \geq t_i\}} \frac{d_i}{n_i(n_i - d_i)} = \frac{\hat{S}(y)(1 - \hat{S}(y))}{n}$$

See wikipedia for proof

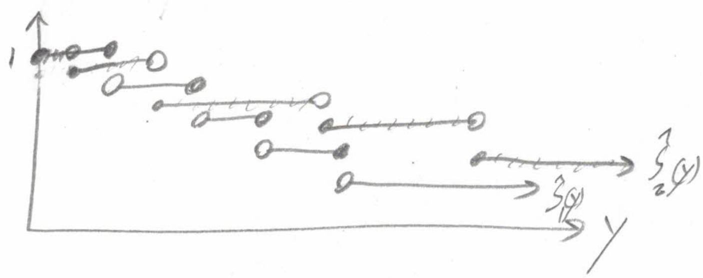
if no censoring

Inference for $\theta = MED(Y)$. Use def of ~~estimator~~ $\hat{\theta}$ and bootstrap like before!

Mean estimation impossible.

Let's say we have two populations and samples $y_{1,1}, \ldots, y_{1,n_1}, c_{1,1}, \ldots, c_{1,n_1}$,
$y_{2,1}, \ldots, y_{2,n_2}, c_{2,1}, \ldots, c_{2,n_2}$

We can use the K-M estimate twice to estimate both survival distr's:



Now we want to prove $H_a: DGP_1 \neq DGP_2$. We can't use K-S anymore as this isn't valid with censoring.

One possibility is to pick a parameter of the curve and compare e.g.

$H_a: Med(Y_1) \neq Med(Y_2) \implies DGP_1 \neq DGP_2$

To test difference in medians, let $\hat{\theta} := \underset{y}{\text{argmin}} \left\{\hat{S}_1(y) \leq \frac{1}{2}\right\} - \underset{y}{\text{argmin}} \left\{\hat{S}_2(y) \leq \frac{1}{2}\right\}$

And then use bootstrap. If $0 \notin CI_{\theta, 1-\alpha} \Rightarrow$ Reject H_0

This difference in medians is usually what you want anyway.

But let's say you really want to prove $H_a: DGP_1 \neq DGP_2$

There is a test called the "Mantel-Cox" test or "Log Rank" test

They proved that:

$$\hat{\theta} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \xrightarrow{d} \chi^2_1$$

Similar to chisq test of independence we saw in 341 class

What is "observed"? It's the sum of the # of observed events in each group $O_1 := \sum_{i=1}^{n} d_{1,i}$, $O_2 := \sum d_{2,i}$

What is "expected"? It is the sum of the expected number of events over all t_i via:

$$E_{1,i} = \underbrace{\frac{n_{1,i}}{n_{1,i}+n_{2,i}}}_{\text{proportion of subjects at risk in group 1}} \underbrace{(d_{1,i} + d_{2,i})}_{\text{\# of deaths across both groups at time } t_i} , \quad E_{2,i} := \frac{n_{2,i}}{n_{1,i}+n_{2,i}} (d_{1,i} + d_{2,i})$$

$E_1 := \sum_{i=1}^{n} E_{1,i}$

$E_2 := \sum_{i=1}^{n} E_{2,i}$

Remember n_i's accounts for censoring

$\Rightarrow E_{1,i} + E_{2,i} = d_{1,i} + d_{2,i}$

It's apportioning the events among the groups in proportion to # of subjects. If this is the same as the # of actual events, there's no difference between the two pop's.